Analyzing the Dynamic Evolution of Hashtags on Twitter: a Language-Based Approach

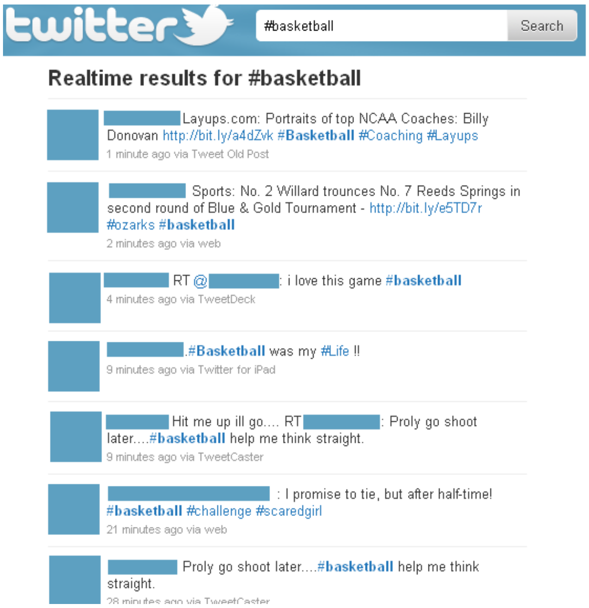
Evandro Cunha[1] **Gabriel Magno**[1] **Giovanni Comarela**[1]
Virgilio Almeida[1] **Marcos André Gonçalves**[1] **Fabrício Benevenuto**[2]
[1]Computer Science Department, Federal University of Minas Gerais (UFMG), Brazil
[2]Computer Science Department, Federal University of Ouro Preto (UFOP), Brazil
{evandrocunha, magno, giovannicomarela, virgilio, mgoncalv, fabricio}@dcc.ufmg.br

Abstract

Hashtags are used in Twitter to classify messages, propagate ideas and also to promote specific topics and people. In this paper, we present a linguistic-inspired study of how these tags are created, used and disseminated by the members of information networks. We study the propagation of hashtags in Twitter grounded on models for the analysis of the spread of linguistic innovations in speech communities, that is, in groups of people whose members linguistically influence each other. Differently from traditional linguistic studies, though, we consider the evolution of terms in a live and rapidly evolving stream of content, which can be analyzed in its entirety. In our experimental results, using a large collection crawled from Twitter, we were able to identify some interesting aspects – similar to those found in studies of (offline) speech – that led us to believe that hashtags may effectively serve as models for characterizing the propagation of linguistic forms, including: (1) the existence of a "preferential attachment process", that makes the few most common terms ever more popular, and (2) the relationship between the length of a tag and its frequency of use. The understanding of formation patterns of successful hashtags in Twitter can be useful to increase the effectiveness of real-time streaming search algorithms.

1 Introduction

The use of hashtags is a way to categorize messages posted on Twitter, an important social networking and microblogging service with 175 million registered users (Twitter, 2010), according to the topic of the message. They can be used not only to add context and metadata to the posts, but also for promotion and publicity. By simply adding a hash symbol (#) before a string of letters, numerical digits or underscore signs (_), it is possible to tag a message, helping other users to find tweets that have a common topic. Hashtags allow users to create communities of people interested in the same topic by making it easier for them to find and share information related to it (Kricfalusi, 2009). Figure 1 shows an example of query for the tag "#basketball", which returns the newest tweets with this hashtag.



Realtime results for #basketball

[redacted] [redacted] Layups.com: Portraits of top NCAA Coaches: Billy Donovan http://bit.ly/a4dZvk #Basketball #Coaching #Layups
1 minute ago via Tweet Old Post

[redacted] [redacted] Sports: No. 2 Willard trounces No. 7 Reeds Springs in second round of Blue & Gold Tournament - http://bit.ly/e5TD7r #ozarks #basketball
2 minutes ago via web

[redacted] [redacted] RT @[redacted]: i love this game #basketball
4 minutes ago via TweetDeck

[redacted] [redacted].#Basketball was my #Life !!
9 minutes ago via Twitter for iPad

[redacted] [redacted] Hit me up ill go.... RT [redacted]: Proly go shoot later....#basketball help me think straight.
9 minutes ago via TweetCaster

[redacted] [redacted] : I promise to tie, but after half-time! #basketball #challenge #scaredgirl
21 minutes ago via web

[redacted] [redacted] Proly go shoot later....#basketball help me think straight.
28 minutes ago via TweetCaster

Figure 1. Example of query for a hashtag on Twitter. Hashtags are not case-sensitive, thus "#basketball" also returns "#Basketball", for example. Tweets with the term "basketball" (without the hash symbol) do not appear in a search for hashtags.

As hashtags are created by the users themselves, a new social event can lead to the simultaneous emergence of several different tags, each one generated by a different user. They can either be accepted by other members of the network or not. In this manner, some propagate and thrive, while others die immediately after birth and are restricted to a few messages.

Similarly, lexical innovations occur when new terms are added to the lexicon of a language, either through the creation of new words, the reuse of existing words or the loan from other languages, for example. An innovation tends to come from one speaker, who proposes it to other members of his speech community – i.e., to whom he is connected in a network of linguistic contacts and influences. Afterwards, these speakers make a cultural selection of the innovation, accepting it or rejecting it.

In the context of the network theory, Figure 2 indicates two moments of a novelty's propagation process: the precise time of the innovation (left) and a later point (right), when some individuals have accepted the innovation, while others, although possibly knowing it, didn't. An innovative linguistic form can get, for some reason, some prestige, and maybe speakers begin to use it, taking it under certain circumstances and transforming it into a variation of the previously hegemonic form.

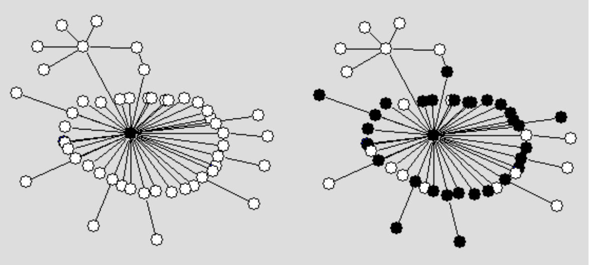


Figure 2. Subgraphs from our Twitter dataset showing two distinct moments in the process of spreading an innovation. The black nodes indicate individuals who joined the innovation (in this case, the hashtag #musicmonday) at a given moment; the white ones indicate individuals who didn't. The links represent follower relationship.

The diffusion of innovations, be they linguistic, behavioral, technological, etc., occurs through a cascade in which the network members, consciously or not, make choices, taking into account a number of factors that determine which forms, behaviors or technologies are more advantageous to be adopted in a given moment (Easley and Kleinberg, 2010).

An important question in the field of linguistics is: how can an initially rare variant spread to an entire linguistic network, or speech community (Sapir, 1921)? How does the linguistic change take out (Silva, 2006)? This change, consisting in the dissemination of less common variants to much of the network or even across the entire network, can be seen as an unexpected fact. However, it occurs. Thus, to better understand the phenomenon of language change, it seems essential to understand the propagation behavior of innovative forms. Understanding how these forms spread – how and where they are born, who are the major disseminators, which network features allow greater dissemination – is the main objective of our research group.

In this work, we examine aspects of the dissemination of hashtags in Twitter, aiming at understanding the process of propagation of innovative hashtags in light of linguistic theories. The utilization of an online social network's dataset allows the review of a linguistic system in its entirety, thereby eliminating the need to work with sampling. It also allows the verification of temporal propagation, enabling a more precise understanding of the path followed by innovations in the network.

Here, we seek to answer mostly two questions: (1) does the distribution of the hashtags in frequency rankings follow some pattern, as the words in the lexicon of a language? (2) Is the length of a hashtag a factor that influences to its success or failure? Our assumption is that identifying linguistic features related to the creation and usage of hashtags in Twitter may raise awareness about individuals' tagging behavior over networks, which is an interesting topic in the field of Network Sciences, Sociology and Social Psychology. Beyond that, this kind of analysis should be interesting to optimize tag recommendation systems not only on Twitter, but on many other online environments, and to increase the effectiveness of real-time streaming search algorithms.

In the next section, we will discuss related works in Linguistics and in Computer Science. We try to always keep contact with linguistic theories, as we believe that complex issues, involving many aspects together, can be better analyzed through a multidisciplinary approach. The following sections cover discussions and the empirical research that was conducted during this study.

2 Related work

Much has been written about linguistic innovations, language variation and language change since Weinreich et al. (1968), which is considered one of the ground works for sociolinguistics. More recently, Troutman et al. (2008) conducted a study with the purpose of simulating language change in a speech community. They built a computational model based on characteristics from language users and from social network structures and tested it in different scenarios, obtaining a probabilistic model that captures many of the key features of language change. Our work extends the traditional way of conducting research on sociolinguistics as we used a corpus of non-natural language data and even so we found compatible results to the ones obtained from natural language data.

Kwak et al. (2010) were the first to study in a quantitative way the topological characteristics of Twitter, information diffusion on it and its power as a new medium of information sharing. Their analyses are in some way related to the ones we perform here. Chew and Eysenbach (2010) led a study that investigated the keywords “swine flu” and “H1N1” on Twitter during the 2009 H1N1 pandemic. The goals of this work were to monitor the use of these terms over time, to conduct a content analysis of tweets and to validate Twitter as a trend-tracking tool. They found the existence of variability in the use of the terms, which is a constitutive aspect of human language. Our findings complement, with more focus on the linguistic approach, what they have discovered, revealing new aspects that can link the creation of hashtags to linguistic innovations.

Romero et al. (2011) studied the mechanics of information diffusion on Twitter. They analyzed the phenomenon of the spread of hashtags, but focusing on the variations of the diffusion features across different topics. Their work introduces the measures “stickiness” – the probability of adoption of one hashtag based on the number of exposures – and “persistence” – which captures how rapidly the influence curve decays. We analyze hashtags as well, but in a different perspective, concentrating on the characteristics that they may have in common with natural language.

3 Dataset and methodology

In our study, we use a dataset consisting of about 2 billion follow links among almost 55 million users. Twitter allowed the collection of data for each existing user, including their social connections, and all the tweets they ever posted. Out of all users, about 8% of the profiles were set private by the users themselves, and only authorized followers could view their tweets. We ignore these users in our analysis. In total, we analyzed more than 1.7 billion tweets posted between July 2006 and August 2009. For a comprehensive description of the data collected we refer the reader to Cha et al. (2010).

As, in some of our analysis, we intend to compare features of the variation of hashtags to linguistic variation, we must find interchangeable hashtags, i.e., different tags used with the same purpose, to characterize messages on the same topic. This corresponds to the basic feature of variant linguistic forms, which are used by different speakers, or at different moments, to name the same object, action etc. Aiming to find interchangeable hashtags, we collected tweets on specific topics. In this way, we could verify the existence of different hashtags used to categorize messages that could be grouped into one category. For example, hashtags like #michaeljackson #mj, #jackson, among many others, refer to the same subject and in a managed environment they would probably be condensed under only one tag.

We selected three relevant topics of this period, namely: Michael Jackson (the singer’s death has been widely reported in the social networks), Swine Flu (the epidemic of H1N1 was a major issue of 2009), and Music Monday (this topic is related to a very successful campaign in favor of posting tweets related to music on Mondays). Then, we built one minor base for each one of the topics: MJ (referring to Michael Jackson), SF (referring to Swine Flu) and MM (referring to Music Monday). These bases were formed by filtering tweets that: (1) included at least one hashtag and (2) included at least one of the following terms that we considered related to the topics: “michael jackson” (for the base MJ), “swine flu” or “#swineflu” (for the base SF), and “#musicmonday” (for the base MM). Consequently, in the base MJ, for example, we gathered all the tweets that included the term “michael jackson” and that had at least one hashtag, even if this tag had no direct relationship with the topic.

Table 1 presents data from each base: number of tweets posted, number of users who posted tweets, number of follow links among users of the base and number of different hashtags used in the tweets of the base.

Base	Tweets	Users	Follow links	Different hashtags
MJ	221,128	91,176	3,171,118	19,679
SF	295,333	83,211	5,806,407	17,196
MM	835,883	196,411	7,136,213	16,005

Table 1. Summary information about the bases built.

4 Comparing Twitter to a natural linguistic system

The directionality of both networks we are studying, i.e. Twitter and speech communities, in addition to the resemblance between the creation of hashtags and linguistic innovations, is an important similarity between these systems. It led to the hypothesis that these structures would have more issues in common.

In this section, we discuss these qualitative similarities, in order to justify the following quantitative comparisons.

4.1 Hashtags and linguistic innovations

A linguistic innovation can be described as any change in any existing language system (Breivik and Jahr, 1989). In linguistics, to say that there was an innovation means that there was a modification, a transformation, in any part of the language – phonetics, phonology, syntax, semantics etc. This novelty is neither degeneration, nor an improvement: language changes and evolves, as a living being, in order to adapt itself to the society in which it is inserted.

We use linguistic knowledge to analyze and explain phenomena related to the creation, usage and dissemination of hashtags. We see similarities between these two systems: like linguistic innovations, new hashtags are created by individuals when they feel the need to categorize their messages with a term not yet used for this purpose. This reflects the speaker’s need to create a term, for example, to name an object or an action that he/she was not acquainted with in the offline world.

Just like hashtags can fail and be used only once, a linguistic innovation may not exceed the boundaries of its creator’s language. An innovation can be used in a specific situation and fall into oblivion, like many linguistic forms which are lost without even being recorded.

4.2 Directionality of the graphs

Twitter’s network can be described as a directed graph. On this social network, relations between users are not necessarily symmetrical, which means that it is possible for someone to follow another person without being followed by him/her. This is very clear when we talk about celebrities who have millions of followers, but at the same time follow only a few users.

This characteristic corresponds to the general absence of directionality of offline social networks. Not only on Twitter the edges can go one-way: in the “real world”, we are somehow connected to celebrities, athletes and famous politicians, and we hear what they say. We are all part of the same speech community, in the sense that a celebrity is able to influence the way we use language. However, they certainly do not even know who we are: it is like on Twitter’s graph, where we follow them, but they do not follow us.

5 Rich-get-richer phenomenon and Zipf’s law

Easley and Kleinberg (2010) characterize what is known as “rich-get-richer phenomenon” or “preferential attachment process”: in some systems, the popularity of the most common items tends to increase faster than the popularity of the less common ones. It generates a further spread of the forms that achieve a certain prestige.

Zipf (1949) examined and confirmed that the frequency of words in English and in other languages follow a power law. Aiming to verify if any kind of pattern is followed in the tags distribution, we analyzed our data from Twitter.

Tables 2 and 3 display information on the distribution of hashtags in each of the bases studied. By “*i*-tweet hashtags”, we mean the hashtags that appear in at most *i* tweets. They are the less common ones. By “*j*-tweet hashtags”, we mean the hashtags that appear in at least *j* tweets, that is, the most popular ones.

Base	% of *i*-tweet hashtags inside the base		
	***i*=1**	***i*=2**	***i*=10**
MJ	59%	72%	88%
SF	59%	73%	92%
MM	60%	74%	91%

Table 2. Distribution of less common hashtags of each base.

Base	number of *j*-tweet hashtags inside the base		
	***j*=10,000**	***j*=5,000**	***j*=1,000**
MJ	3	6	28
SF	3	4	14
MM	2	3	28

Table 3. Distribution of most popular hashtags of each base.

The percentage of hashtags according to the number of tweets in which they appear are remarkably very similar in the three bases. It seems to confirm the possible existence of a “rich-get-richer” pattern: few hashtags – the most popular ones – are used in most of the tweets, while the vast majority of them are used in only a few posts. Table 2 shows that around 60% of hashtags are used only once in tweets of the respective base, i.e. do not propagate to the rest of the network; around 90% of them are not used more than ten times, which shows that the great part of the hashtags get restricted to only one user or to a very small community of users.

On the other hand, just like Zipf (1949) showed for natural languages, the most used hashtags get very high frequencies of use. Table 4 shows data from the three most used hashtags in each of the bases and makes clear that, also on Twitter, a person´s behavior depends on the choices made by other people (Easley and Kleinberg, 2010).

Complementing these data, Figure 3 associates the position of a hashtag in a popularity ranking (based on the number of times that a hashtag has been used) to the volume of tweets in which it appears. A plot in log-log coordinates, where *x* is a rank of a tag in the frequency table and *y* is the total number of the tag’s occurrences in tweets, shows that the distribution of hashtags on Twitter also follow the general trend of a Zipfian distribution, appearing approximately linear on log-log plot.

Base	Most used	2nd most used	3rd most used
MJ	#michaeljackson 35,861 12.3%	#michael 27,298 9.3%	#mj 16,758 5.7%
SF	#swineflu 230,457 51.5%	#h1n1 70,693 15.8%	#swine 12,444 2.8%
MM	#musicmonday 824,778 79.7%	#musicmondays 11,770 1.1%	#music 5,106 0.5%

Table 4. Data from the most used hashtags of each base. Below each hashtag are given the number of times it was used and the percentage that it represents of the total use of hashtags in the base.

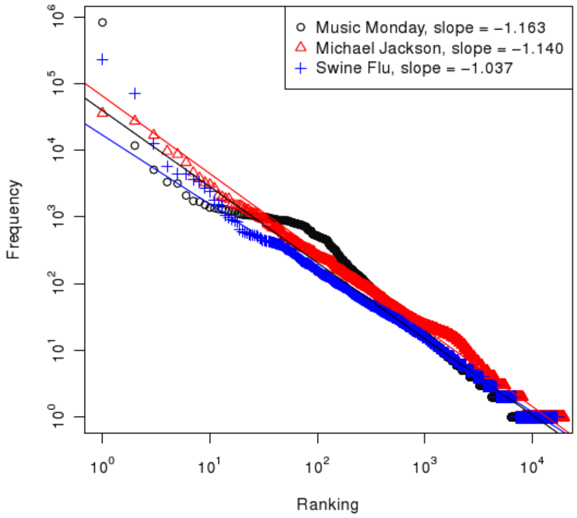


Figure 3. A log-log plot showing volume of tweets in which the hashtag was used vs. its position in a popularity ranking.

Only three values on the left, which refer to tags that occupy the top positions in the frequency ranking (and thus were used more often), are not well described by the interpolations: the most frequent tag on MM base and the two most frequent ones on SF base. This is due to the very high usage of these hashtags: #musicmonday appeared in almost 830,000 tweets of its base; #swineflu, in more than 230,000; and #h1n1, in more than 70,000. The other values, however, show that this is a very good fitting model for our purposes.

It is interesting to notice the similarity of results despite being completely different topics. Even the slopes of the interpolation curves are similar, varying from -1.037 to -1.163.

6 Hashtag length and frequency

Each word or phrase spoken by someone tells a story and reflects characteristics of this individual and his/her group. According to the Theory of Language Variation and Change (Weinreich et al., 1968; Labov, 1995, 2001), lexical choice is the result of a series of social interactions that make up and form, little by little, the individual speech. Naturally, these interactions and influences are so subtle that we ourselves hardly realize them: gender, age, location, social role, hierarchical position in an organization – all this reflects the way we use language in various situations of everyday life. Understanding what makes speakers choose one of the forms in variation, in certain situations, is one of the goals of Sociolinguistics.

In addition to these social factors that influence the way we express ourselves, described by Labov (2001), there are also many strictly linguistic factors which perform such influence, as Labov (1995) presents. One of these factors seems to be the length of the words, as noted by Zipf (1935) and analyzed by Sigurd et al. (2004).

Zipf (1935) suggests that the length of a word tends to bear an inverse relationship, not necessarily proportionate, to its relative frequency. Sigurd et al. (2004) analyze data from different text genres in English and Swedish and corroborate the hypothesis, showing that longer words tend to be avoided, presumably because they are uneconomic.

Given this evidence, and considering the concern of Twitter users to save space, since the maximum size of each tweet is 140 characters, we investigate whether the length of a hashtag is one of the strictly linguistic factors that influence on their success or failure.

In order to carry out this analysis, we compared the length of the most popular hashtags in each of the bases with the less popular ones. We noticed that the most popular ones are simple, direct and short; on the other hand, among those with little utilization, many are formed by long strings of characters. Table 5 displays preliminary information about the length of hashtags and popularity and shows that hashtags formed by 15 or more characters are not present among the most used tags.

Table 6 lists the average length, in number of characters, of different groups of hashtags, divided according to their positions in the ranking of frequency of each base.

Most common hashtags (number of tweets)	**Most common hashtags with 15 or more characters (number of tweets)**
#michaeljackson (35,861) #michael (27,298) #mj (16,758)	#nothingpersonal (962) #iwillneverforget (912) #thankyoumichael (690)
#swineflu (230,457) #h1n1 (70,693) #swine (12,444)	#swinefluhatesyou (1,056) #crapnamesforpubs (145) #superhappyfunflu (124)
#musicmonday (824,778) #musicmondays (11,770) #music (5,106)	#musicmondayhttp (540) #fatpeoplearesexier (471) #crapurbanlegends (23)

Table 5. Confrontation of most common hashtags and most common 15-character hashtags. In front of each hashtag is given the number of times it was used in tweets of the base.

Topic	**Average length of...**					
	...the *k* most popular hashtags					**...the less popular hashtags**
	***k*=10**	***k*=20**	***k*=30**	***k*=40**	***k*=50**	
MJ	7.1	6.85	7.8	8.02	7.74	10.16
SF	5.3	7.35	7.17	7.2	7.04	10.3
MM	9.5	8.4	7.27	6.4	5.92	11.66

Table 6. Average length of the most and the less popular hashtags. The samples with the less popular hashtags were formed by 50 randomly selected hashtags among those which appeared only in one tweet of each base.

In all of the bases, the average length of the most popular hashtags is considerably lower to the average length of the less popular ones. Figure 4 compares data from Table 6, including information about standard deviation. It is clear that the differences between the lengths of the few most popular tags are not relevant, as the average lengths of the *k* most popular tags, with *k*={10,20,30,40,50}, are roughly similar and do not follow a fixed pattern. However, the comparison with 1-tweet hashtags (less popular ones) shows important differences which led us to believe that the length of a hashtag may be an internal factor – or a strictly linguistic factor – that determines the success or the failure of tags on Twitter, even if more accurate study is needed at this point.

This reflects the small number of hashtags composed of complete sentences (such as #mileycometobrazil, #herewegoagain and many others) occupying good positions in the popularity rankings. Their low standard of success can be attributed to some reasons besides their increased length, such as: (1) sentences admit high rate of variation (e.g. #thankyoumichael, #thanksmj, #michaeljacksonthanks), which reduces the frequency of each of the competing forms; (2) sentences are more difficult to memorize, as they may accept different word orders; and (3) in sentences, it seems to be more prone to misspellings (as in #thanktyoumichael), maybe because of the apparent difficulty of reading the terms without the ordinary spaces between them (we believe that it is easier to notice the misspelling in "thankt you michael" than in "thanktyoumichael", though this is an assumption that must be verified through more extensive work in Psycholinguistics and Applied Linguistics).

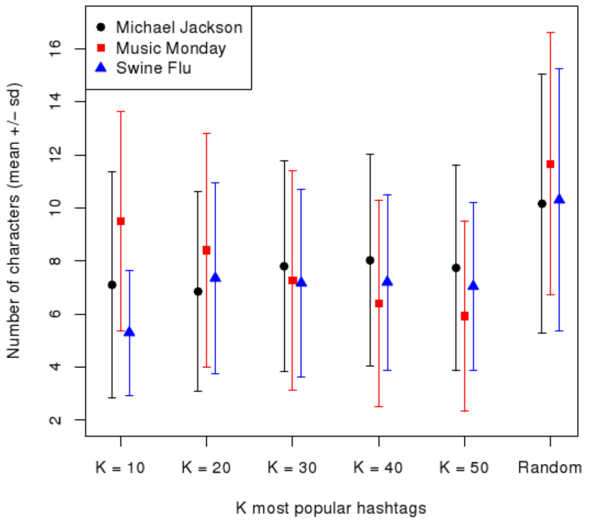


Figure 4. Average number of characters of the most popular hashtags and of a randomly selected sample of 50 less common tags.

7 Underscores in hashtags

We conducted an analysis to check the influence of the only sign allowed in the formation of hashtags besides letters and numbers: the underscore (_). In all the bases, the use of the sign _ led the hashtags to low popularity rankings: #michael_jackson reached position 248 in its base, with only 128 tweets; #swine_flu reached position 67 in its base, with no more than 246 tweets; #music_monday wasn't even used. Table 7 shows the use of sign _ in hashtags. Here, we call a "_-hashtag" any hashtag in which has been used the sign _.

Base	**Number of _-hashtags**	**% of _-hashtags among *i*-tweet hashtags**	
		***i*=2**	***i*=10**
MJ	251 (1.2%)	89%	97%
SF	155 (0.9%)	87%	97%
MM	143 (0.9%)	89%	98%

Table 7. Distribution of hashtags containing the sign "_".

We can observe that almost all of the _-hashtags have lower positions in the popularity rankings: at least 97% of them are used in 10 or less tweets, which seems to indicate rejection to this sign. Once again, the distributions corresponding to each of the bases are similar, suggesting a uniform behavior across the whole network.

8 Conclusion

This paper examines, through a language-based approach, some issues concerning the formation and the usage of hashtags on Twitter. We proposed that linguistic theory could be used to formulate hypothesis on online systems like Twitter and our analysis showed not only qualitative, but also quantitative similarities between offline and online speech communities.

We revealed interesting aspects about the distribution of hashtags according to their popularity, associating it to the distribution of words in frequency rankings. We also went further on the question suggested by Romero et al. (2011), who proposed to consider what distinguishes a hashtag that spreads widely from one that fails to attract attention: we could find that the tag's length, for example, is one of these factors. This kind of analysis can be a useful tool for tag recommendation systems in different environments, but there are a number of other aspects which can be considered in future work and that can collaborate to the study of human tagging behavior.

References

Breivik, L.E., and Jahr, E.H. (Eds.) 1989. *Language change: Contributions to the study of its causes*. Berlin/New York: Mouton de Gruyter.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. (2010). Measuring user influence in Twitter: The million follower fallacy. *Int'l AAAI Conference on Weblogs and Social Media (ICWSM'10)*. Washington DC, USA.

Chew C., and Eysenbach G. 2010. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE 5(11)*: *e14118*. doi: 10.1371/journal.pone.0014118

Easley, D., and Kleinberg, J. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge: Cambridge University Press.

Kricfalusi, E. 2009. The Twitter hash tag: What is it and how do you use it? Retrieved from http://tinyurl.com/bw85z2

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a social network or a news media? *International World Wide Web Conference (WWW 2010)*. Raleigh, USA.

Labov, W. 1995. *Principles of linguistic change: Internal factors*. Reprint. Oxford/Cambridge: Blackwell.

Labov, W. 2001. *Principles of linguistic change: Social factors*. Oxford/Cambridge: Blackwell.

Romero, D., Meeder, B., and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. *International World Wide Web Conference (WWW 2011)*. Hyderabad, India.

Sapir, E. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace and World.

Sigurd, B., Eeg-Olofsson M., and Van de Weijer, J. 2004. World length, sentence length and frequency – Zipf revisited. *Studia Linguistica 58(1)*, (pp.37-52). Oxford/Malden: Blackwell.

Silva, L.G. 2006. A dimensão sociolingüística do Atlas Lingüístico do Brasil. *Anais da VIII Semana de Letras da Universidade Federal de Ouro Preto*. Ouro Preto, Brazil: Universidade Federal de Ouro Preto.

Troutman, C., Clark, B., and Goldrick, M. 2008. Social networks and intraspeaker variation during periods of language change. *Proceedings of the 31st Annual Penn Linguistics Colloquium*. (pp.325-338). Philadelphia: University of Pennsylvania.

Twitter, 2010. About Twitter: A few Twitter facts. Retrieved from http://twitter.com/about.

Weinreich, U., Labov, W., and Herzog, M. 1968. Empirical foundations for a theory of language change. In Lehmann W., and Malkiel Y. (Eds.), *Directions for historical linguistics* (pp.97-195). Austin: University of Texas Press.

Zipf, G.K. 1935 (reprinted 1965). *The psycho-biology of language*. Cambridge: MIT Press.

Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.